对等传输流量优化技术研究

张国强 唐明董 程苏琦 张国清

摘　要： 对等传输（P2P）内容分发系统能大幅减轻对等传输内容提供商的分发代价，提高系统的可扩展性。但是，覆盖网络和底层网络的不匹配导致了网络资源的浪费，产生了大量的冗余流量，激化了内容提供商和网络运营商（ISP）之间的矛盾。因此，如何有效利用底层网络的带宽资源，降低对等传输内容分发系统对 ISP 网络的流量压力，是目前对等传输内容分发系统研究的热点，也是对等传输系统可持续发展的关键。本文从对等传输缓存、位置感知技术和数据调度算法三个维度出发详细阐述了不同的对等传输流量优化技术，探讨了各种技术的适用环境，指出了存在的问题和未来的方向。

关键词： 对等传输流量本地化、对等传输流量优化、对等传输缓存、位置感知、网络编码、网络拓扑

1　引言

近年来，基于对等传输（peer-to-peer，P2P)模式的内容分发系统获得了用户广泛的青睐，如 Kazaa、Gnutella、Emule、BitTorrent、PPLive、PPStream 等。较之传统的客户/服务器模式或内容分发网络（CDN）模式，对等传输内容分发模式大大降低了内容提供商的分发代价，增强了系统的可扩展性。

但是，对等传输应用产生的巨大流量成为了制约其进一步发展的瓶颈。测量表明，对等传输流量已经超越 Web，成为网络流量的最主要贡献者。目前，在终端用户、对等传输内容提供商和网络运营商（ISP）构成的三角中，前两者都从对等传输应用的普及中获益良多。终端用户提高了下载的速度，内容提供商降低了分发代价。网络运营商则成为了唯一的受害者。网络运营商为网络带宽所做的扩容被对等传输应用无情地消耗，导致投资无法获得收益。此外，对带宽的过度占用影响了其它业务的正常运行。虽然网络运营商可以通过提高用户的资费来平衡其投入/收益比，但这又面临着流失用户的潜在危险。

对等传输产生巨大流量的一个原因在于没有有效地利用底层的资源，造成了大量的资源浪费。因此，优化对等传输的流量成为缓解、消除内容提供商和网络运营商之间紧张关系的唯一途径。对等传输的流量包括控制流和数据流两部分，其中，实际的数据传输居对等传输网络流量的主导地位。因此，优化实际的数据传输是关键。

内容提供商和网络运营商为解决对等传输产生的巨大流量问题所做的努力大致可分为三个阶段：对抗、各自为战和合作共赢。在对抗阶段，网络运营商采取深度包检测（DPI, Deep Packet Inspection）等技术对对等传输流量进行识别，从而对对等传输流量进行封堵、限速、整形。作为应对，对等传输内容提供商利用动态端口、消息加密等机制来隐藏流量。这导致了类似病毒和反病毒的恶性循环，业界戏称其为双方的一场“战争”。在第二阶段，双方都采取了一些积极措施来优化网络资源的使用。内容提供商依据反向工程的方法来推测底层的拓扑结构和网络状态信息，并依据这些信息来辅助建立与底层网络匹配的覆盖网拓扑，从而提高带宽资源的利用率。网络运营商则通过对对等传输流量进行缓存，或通过 Proxy-Tracker（代理-追踪器）等技术来影响对等传输覆盖网拓扑的建立，达到优化网络资源利用的目的。

第三阶段，内容提供商和网络运营商开始尝试以合作的方式来优化网络资源的使用，即由网络运营商提供网络拓扑和状态信息的服务，而内容提供商通过访问这些服务来优化其拓扑建立和数据调度。在该架构下，网络运营商可以控制提供的网络拓扑和状态信息的粒度，隐藏必要的私密信息，并可通过提供服务获得收益。而内容提供商可以通过该服务获得准确的位置信息，从而优化其资源调度，减少了拓扑探测的开销，也避免了被网络运营商封堵的危险。

对等传输流量优化主要包括三类技术：对等传输缓存技术、位置和拓扑感知技术、数据调度算法。本文将以此为出发点，深入剖析和比较这三类技术。在详细介绍这三类技术之前，首先对对等传输内容分发系统进行简单的分类。

2 对等传输内容分发系统分类

按所分发内容的性质，对等传输内容分发系统可分为非实时文件分发系统(如 Gnutella、Kazaa、BitTorrent)和实时流媒体分发系统(如 PPLive、PPStream)。按对等传输网络的结构，则可分为非结构化的对等传输和结构化的对等传输。结构化对等传输建立在分布式哈希(DHT)技术之上，其优点是搜索代价低，缺点是网络的维护代价高，也不能有效支持基于关键词的模糊查询，因此在内容分发系统中并没有获得广泛的应用。目前对等传输系统的主要流量都来自于非结构化对等传输应用。在非结构化的对等传输中，按照所采用的技术和出现时间先后，又可以大致分为三代。Napster 是第一代对等传输内容分发系统的典型代表，它包含一个集中的索引服务器；第二代包括完全分布式的 Gnutella, 以及混合式的 Gnutella 和 Kazaa 等；第三代包括文件分发系统 BitTorrent，实时流媒体播放系统 PPLive 等。然而，虽然都采用了对等传输技术，Gnutella, BitTorrent 和 PPLive 也有着显著的区别。BT 和 PPLive 与 Kazaa 的最大区别在于：在 BT 或 PPLive 中，一个节点只需下载整个文件的一部分就可以向其它节点提供服务。其覆盖网络既是控制信息的载体，也是实际数据传递的载体。而在 Kazaa 中，只有拥有完整文件数据的节点才能向其它节点提供下载服务。即，覆盖网只作为控制信息的载体，而实际的数据下载并不依赖于覆盖网络，下载过程亦不存在合作网络的概念。后面我们将看到，不同的对等传输流量优化技术对这两种不同的对等传输文件分发模式有不同的适用性。

3 对等传输缓存技术

缓存技术是网络运营商用于缓解流量压力、缩短用户响应时间的一种方案，最早被广泛用于 Web 服务。随着对等传输流量的激增，对其进行缓存成为网络运营商缓解流量压力的首选。但是，对等传输与 Web 具有不同的流量特征，而且两者缓存设计的目标也不同。网页（web）缓存设计的一个主要目标是缩短用户感知的响应时间，而对等传输缓存的主要设计目标是降低网络流量。网页缓存一般以对象命中率(hit rate)和用户的平均响应时间作为缓存算法的评价指标，而对等传输缓存则一般以字节命中率(byte hit rate)作为评价指标。因此，应根据对等传输的流量特征和缓存设计目标为对等传输应用量身定制高效的缓存算法。

3.1 对等传输流量特征

对等传输应用的流量特性与传统的 Web 流量存在显著的差别，如表 1 所示。

两种应用流量特征的不同对缓存设计的不同方面有着不同程度的影响。例如对象大小、对象流行度是影响缓存效率的主要因素，而协议开放性则影响缓存的实现复杂度和可扩展性。

表1. 对等传输流量和 Web 流量的主要区别

	对等传输	Web
对象大小	大致包含三个级别的负载：10M 以下的小文件，几百兆的中等文件，大于 1G 的大型文件	一般都比较小
对象流行度	流行度不服从齐普夫定律[1]，最流行的对象的流行度远远低于齐普夫定律的预测结果	流行度服从齐普夫定律
对象流行度变化	流行度会发生突变。对象可能会在一夜之间变得流行，并在短时间内流行度迅速下降	对象的流行度变化比较平稳
对象可变性	对象不可变	越来越多的可变对象
用户获取次数	大部分对象用户只获取一次	同一对象用户可以多次访问
下载一个对象的会话数	可以同时开启数十个会话	一个或少数几个会话
会话持续时间	可能持续数小时	会话通常在几秒内完成
协议开放性	大量的私有协议	基于 HTTP 的标准协议
端口	不同的网络有不同的端口	单个端口（80）

3.2 对等传输缓存算法

缓存算法包含两个核心问题：(1)缓存全部对象还是缓存部分对象？(2)缓存替换算法，即当缓存空间不够时，如何替换缓存对象？

在网页缓存中，由于对象通常都较小，因此一般都缓存整个对象。而在对等传输系统中，由于对象可能非常大，如果缓存整个对象，那么一个缓存中能够缓存的对象个数将非常有限，这可能会降低缓存的效果。因此，将对象分块或分段，让缓存系统仅缓存对象的一部分就成为很好的选择。

缓存替换算法是决定缓存效率的另一个主要因素。传统的缓存替换算法包括 LRU(最久未使用)、LFU(最近最少使用)、MINS(最小对象大小)、MAXS(最大对象大小)等。但这些算法无法很好地适用于对等传输系统。LSB(Least-Sent-Byte)[2]是一种针对对等传输的缓存替换算法，在进行对象替换时，总是替换给用户提供的字节数最少的那些对象。

对等传输系统中对象的流行度和流行时间与传统的网页对象不同，这也对设计缓存具有指导意义。调查表明，对等传输对象的流行度会在短时间内达到顶峰，随后开始迅速下降，在 5-10 周内流行度会降低到 1/6[3]。这意味着，如果采用 LFU 等根据访问频率的缓存替换算法，这类对象会在它们流行度降低后的很长一段时间继续呆在缓存中，导致缓存效率低下。另一方面，对象的流行时间并非十分短暂，通常在 2-3 个月左右。这表明可以通过缓存部分对象的方法逐步增加对象被缓存的比例，使缓存系统有充足的时间来建立对象流行度的档案信息。这样，缓存系统就可以识别出真正流行的对象，并逐步增加其缓存的比例直到整个对象都被缓存为止。

3.3 对等传输缓存技术存在的问题

虽然对等传输缓存技术是网络运营商目前进行对等传输流量优化的首选方案，但它也存

[1] 齐普夫（Zipf）定律是美国学者 G.K.齐普夫于本世纪 40 年代提出的词频分布定律。它可以表述为：如果把一篇较长文章中每个词出现的频次统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然数给这些词编上等级序号，即频次最高的词等级为 1，频次次之的等级为 2，……，频次最小的词等级为 D。若用 f 表示频次，r 表示等级序号，则有 f×r＝C(C 为常数)。符合这一定律的概率分布成为齐普夫分布。

在下述问题：

1. 有效性

只有当对等传输节点的位置感知能力不强时，对等传输缓存技术的效果才会较好。但随着对等传输节点位置感知能力的提高，对等传输缓存的有效性将相应地下降。

2. 可扩展性

由于每种对等传输系统都采用不同的协议，有些甚至是不公开的私有协议，限制了对等传输缓存系统的扩展性。为了对一种新的对等传输应用系统的流量进行缓存，必须首先了解该对等传输应用的协议细节，这增加了实现的复杂性，降低了系统的扩展性。与之相较，Web 应用基于开放的协议，使得网页缓存具有通用性，实现要容易得多。

3. 版权问题

虽然网页缓存也存在版权问题，但由于对等传输内容分发系统分发对象的版权敏感性更强，因此版权问题显得尤为突出。对等传输缓存系统应小心避免触犯任何版权保护法例。需要注意的是，在中国至今仍然没有明确的相关法律法规。在全球范围内对相关问题有明确界定的也仅仅是美国的《千禧年数字版权法(DMCA)》，因此在缓存系统的数字版权保护方面，留下了极大的灰色地带[4]。

4. 利益格局问题

对等传输缓存违背了整个网络的运营模式，网络运营商部署对等传输缓存是不得已而为之的无奈之举，无法从根本上解决网络运营商和对等传输内容提供商之间的矛盾。

4 位置感知技术

对等传输网络是构建于底层网络之上的覆盖网，它与底层网络之间的匹配程度很大程度上决定了对等传输节点对底层网络资源的使用行为。早期的对等传输网络采用随机的方式选择邻居，导致构建出的覆盖网与底层网络拓扑之间存在严重的不匹配。这不仅降低了应用层的性能，也导致了大量的冗余流量。图 1 是采用位置感知方式(图 1(a))和随机方式(图 1(b))构建出的覆盖网示例。采用随机方式构建出来的覆盖网络与底层网络可能会出现严重的不匹配，因而在应用层数据传输时会产生很多的长距离传输，浪费宝贵的带宽资源。测量显示[5]，在 Gnutella[2]网络中，只有 2-5%的连接是位于同一自治系统(Autonomous System, AS)内部的，但超过 40%的 Gnutella 节点是位于最大的 10 个自治系统中，这表明 Gnutella 网络基本不具备位置感知能力。利用位置感知技术，可以使构建出的覆盖网络和底层拓扑在路由层具备一致性，如图 1(a)所示，底层网络相近的节点之间连接更为紧密，从而达到优化网络资源的目的。

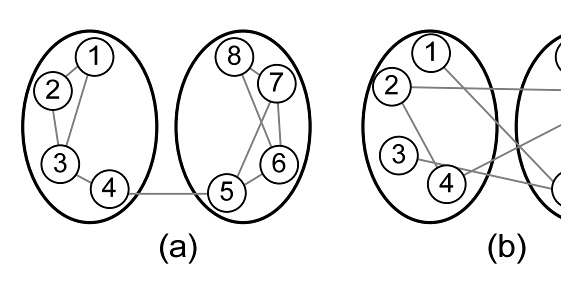


图1. 两种对等传输拓扑构建方式
(a)基于位置感知技术构建的对等传输网络;
(b)基于随机邻居选择构建的对等传输网络

[2] 一种文件共享网络

上述例子表明，构建一个与底层网络拓扑匹配的覆盖网对于改善对等传输系统的性能和降低对网络的流量压力至关重要。这涉及两方面的研究内容：一是如何获取和提供位置感知信息，二是如何利用位置感知信息来辅助拓扑构建。

4.1 位置感知信息的获取

位置感知的目的在于预测节点之间的网络距离或对候选节点的邻近度进行估计。依据所采用技术的实时性，可以分为实时和非实时的方法。

4.1.1 基于实时测量的位置感知技术

根据是否采用网络坐标的形式，基于实时测量的位置感知技术可分为非坐标的位置感知技术和基于网络坐标的位置感知技术。

非坐标的位置感知技术可以依赖新建的专用基础设施(如 IDMaps 和 Binning)，或利用现有的网络基础设施(如域名系统和内容分发网络)来感知节点之间的相对网络距离或邻近度。

IDMaps[6]包括追踪服务器 Tracer 和地址前缀（AP，Address Prefix)两个概念。Tracer 是一组分布在整个 Internet 上的专用测量服务器。Tracer 周期性地测量节点间的时延，同时也测量到与之地址前缀邻近的节点的时延。需要访问 IDMaps 提供的服务的客户端收集所有这些时延信息，构建一张由 Tracer 和地址前缀组成的互联网的虚拟拓扑图(如图 2 所示)，并基于该虚拟拓扑图估算任意两个 IP 地址的时延。具体而言，对给定的两个 IP 地址，其时延估计值为这两个 IP 地址所属的 AP 到各自最近的 Tracer 的时延与这两个 Tracer 的时延之和。

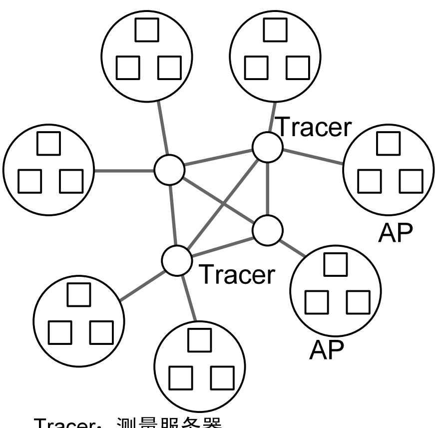


图2. IDMaps 的结构

文献[7]提出的装箱方法(Binning)是一种基于地标节点的测量方法。设地标节点为 *L1, L2,…, Lk*。当节点 P 加入系统时，首先测量与这 *k* 个地标节点的延迟 *l1, l2,…, lk*。对应得到其他节点相对一个地标节点延迟大小顺序的一个排列 *L1', L2', …, Lk'*。该排列本身就可以作为节点的一个分类依据。如果节点 P1 和 P2 经过上述测量和排序得到了相同的地标节点排列，则将 P1 和 P2 归为同一类。这是因为拓扑上相近的节点到不同的地标节点应具有相似的延迟，因而对应的地标节点排序也相同。在地标节点的排序基础上，还可对延迟的绝对值进行等级划分，以获得更细粒度的分类。

由华盛顿大学（University of Washington）开发的工具 King[8]利用现有的域名解析系统（DNS）体系架构进行时延预测，其优点在于无需部署额外的测量平台，也不需要对现有的协议体系做任何修改。King 的具体工作方式如下：任意给定两个节点 A 和 B，首先找到与 A 和 B 邻近的权威名字服务器 NS(A)和 NS(B)，然后依靠域名解析系统的递归查询来获取 NS(A)和 NS(B)之间的时延，并将该时延作为节点 A 和 B 之间的时延估计。图 3 给出了 King 的工作流程示意图。为了避免域名解析服务器从域名体系结构的根服务器开始逐级查询，在测量时延前，需要先进行一次预解析，使得域名服务器 NS(A)得以缓存 NS(B)的地址。但是，并非所有的权威服务器都与目标主机是邻近的，因此该方法可能会产生无法忽视的误差。Turbo King[9]对此进行了改进，降低了误差范围，同时克服了 King 导致的大规模域名解析系统缓存污染问题。

文献[10]提出了一种基于内容分发网络的信息重用方法来预测两个节点的邻近度。内容分发网络使用动态域名解析系统重定向技术为客户提供低延迟的镜像服务器，这使得我们可以通过利用现有内容分发网络基础设施的重定向行为实现较优化的节点选择。这种方案假定任何两个具有相似的重定向行为的节点有很高的概率是邻近的，如：有较大的可能属于同一个网络运营商。

基于网络坐标的位置感知技术的基本原理是将网络中所有节点依据网络距离(可以是延迟、带宽和丢包率等)构成的空间嵌入到一个虚拟的空间，并为网络中的每个节点赋一个虚拟空间中的坐标。之后，任何两个节点之间的网络距离(如往返时间[3])便可根据节点的虚拟坐标(如坐标之间的欧氏距离)进行计算。网络坐标系统的一大优点是消除了网络节点为了确定延迟而需要付出的额外探测开销，增强了网络相对位置感知系统的扩展性。例如美国卡内基梅隆大学开发的 GNP[12]是一种基于静态地标节点的网络坐标系统，另外还有基于动态地标的网络坐标系统，如美国麻省理工学院开发的 Vivald[11]。

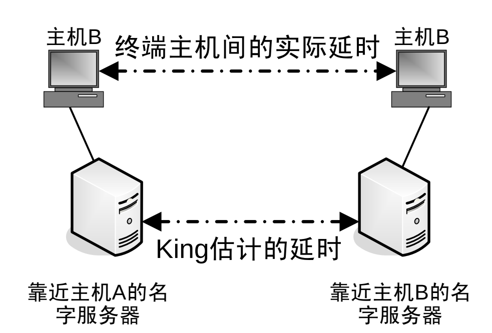


图3. King 依据域名解析系统估计节点延迟的示意图

4.1.2 非实时的位置感知技术

非实时位置感知技术主要是基于 IP 地址结构进行邻近性估计。互联网是基于 IP 地址前缀进行路由。每个 IP 地址都对应了一个或多个可路由的 IP 地址前缀，可表示为 a.b.c.d/l，l 表示前缀长度。但如果没有路由系统的信息支持，节点无法获知一个 IP 地址的确切前缀长度。传统的全球可路由的单播 IP 地址包含 A、B、C 三类，节点可以简单地根据 IP 地址确定所属的类别，从而确定自身的前缀长度为 8、16 或 24。但是由于子网技术和无类域间路由选择（classless inter-domain routing，CIDR）技术的广泛使用，使得上述方法在准确度和粒度上都存在不少问题。

一个更恰当的方案是利用路由系统的信息。路由表保存了从该路由器可见的所有 IP 地址前缀信息。存在两类路由表：用于自治系统域间路由的边界网关协议（Border Gateway Protocol, BGP）路由表和用于自治系统内部的域内路由表。通常，域内路由表对本域的路由条目具有更细的粒度。但网络运营商一般不愿公开自己的域内路由表，因为这被视作隐私信息。相对来说，域间路由表的获取渠道更多。

上述方法的一个缺点是需要依赖于事先收集的路由表前缀信息，因此无法及时对网络地址分配的变化做出调整。文献[13]提出了一种自组织分布化的前缀匹配方法。每个节点将自己 IP 地址的掩码，或是随机的 k 比特哈希成一个键值，然后将自己的 IP 地址存储在分布式哈希表系统中该键值对应的节点上。这样，一个新加入的节点可以很容易找到与自身 IP 具有相同 IP 前缀的节点。该方法不需要依赖预处理的 IP 地址前缀集合，是一个完全自组织和自适应的系统，但需要维护一个分布式哈希表。

一类更粗粒度的邻近度估计方法是将 IP 地址映射到自治系统号码，从而可以判断不同的节点是否位于同一个自治系统。将 IP 映射到自治系统号可以通过 BGP 的路由表实现。BGP 的路由表保存了 IP 地址前缀到达某个自治系统的路径（AS 路径）。AS 路径中的第一个自治

[3] round trip time, RTT

系统被称为 IP 地址前缀的源自治系统。综合多个边界网关协议路由表可以构建出 IP 地址前缀和源自治系统的映射表。给定一个 IP 地址，通过最长前缀匹配即可找到对应的自治系统号码。

4.1.3 基于运营商提供的信息

上述方案都是由内容提供商或第三方通过反向工程的方法来感知网络节点的相对距离。实际上，网络运营商是提供网络拓扑和网络状态信息的最佳者。近来，开始出现了由网络运营商提供位置信息服务的技术提案。但由于这涉及网络运营商和内容提供商的相互协作，因此提供松耦合、可扩展、安全、通用和具有相对灵活性的标准化服务接口是这类方法的基本要求。

P4P[0] 是这类技术的典型代表，它允许应用与网络提供商之间有效地合作，在提高或者保持对等传输的应用层性能的前提下，更有效和公平地利用网络资源。

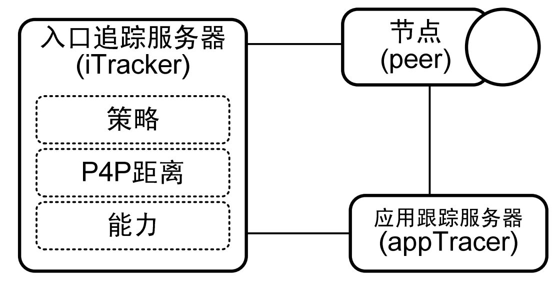


图4. P4P 控制层的实体和信息流

P4P 架构提出在网络运营商网络中部署一种入口追踪服务器 iTracker，这种设备收集网络运营商网络的信息，并提供与对等传输应用交互的接口。这些接口包括：静态网络策略(policy)、反映网络策略和网络状态的 P4P 距离(P4P distances)、网络能力(如缓存等)。这些接口能够支持外部应用获得网络运营商网络的相关信息，同时保护了网络运营商的隐私，从而实现网络运营商和内容提供商联合优化他们各自的网络性能。图 4 给出了一个 P4P 的控制层的实体和信息流描述。其中，appTracker 是对等传输应用的 Tracker（追踪器，用于记录下载名单）设备，它与 iTracker 交互并将 P4P 控制层的信息转发给对等传输应用节点。

目前，已经成立了 P4P 工作组来推动 P4P 的标准化工作。这得到了美国部分网络运营商和内容提供商的欢迎。但是 P4P 仍然面临着挑战，主要是网络运营商和内容提供商是否有足够的动力合作并采用 P4P 体系结构还是个问题。P4P 架构目前也尚未解决如何以一种可扩展和高效的方式实现的问题。

网络匹配服务(Network Matching Service: NMS)是中科院计算所、电信研究院及各主要网络运营商向中国通信标准化协会（CCSA）提出的《基于承载网感知的 P2P 流量优化技术框架》[14]标准的主要技术，目前已经成为国家通信行业报批稿。该框架定义了一种提供信息帮助对等传输应用进行决策的服务，称为网络匹配服务。图 5 是网络匹配服务的基本架构，其中包括了承载网信息提供者、网络匹配服务器、网络匹配服务发现服务器和网络匹配服务客户等功能实体。

网络匹配服务客户可以是一般的对等传输主机或对等传输索引服务器，它被授权访问一些网络匹配服务器。网络匹配服务客户可以利用网络匹配服务实现资源提供节点的优化选择、优化对等传输网络拓扑、优化中继节点或缓存节点的选择等。网络匹配服务器是由网络运营商实现和部署的用于提供网络匹配服务的实体。一个网络运营商可以在自己的网络内自主地部署一个或多个网络匹配服务器。每个网络匹配服务器可以制定策略规定向哪些对等传输应用和哪些对等传输主机授与访问权限。“网络匹配服务发现服务器”的功能是发现网络匹配服务器，提供注册服务。网络运营商将自己的网络匹配服务器的地址和策略注册到网络匹配服务发现服务器上。对等传输主机通过网络匹配服务发现服务器找到对自己授权的网络

匹配服务器。实现网络匹配服务所需的网络拓扑、网络策略、网络能力、网络动态等信息通过承载网提供，这里所说的“承载网”特指网络运营商或它的代理程序。

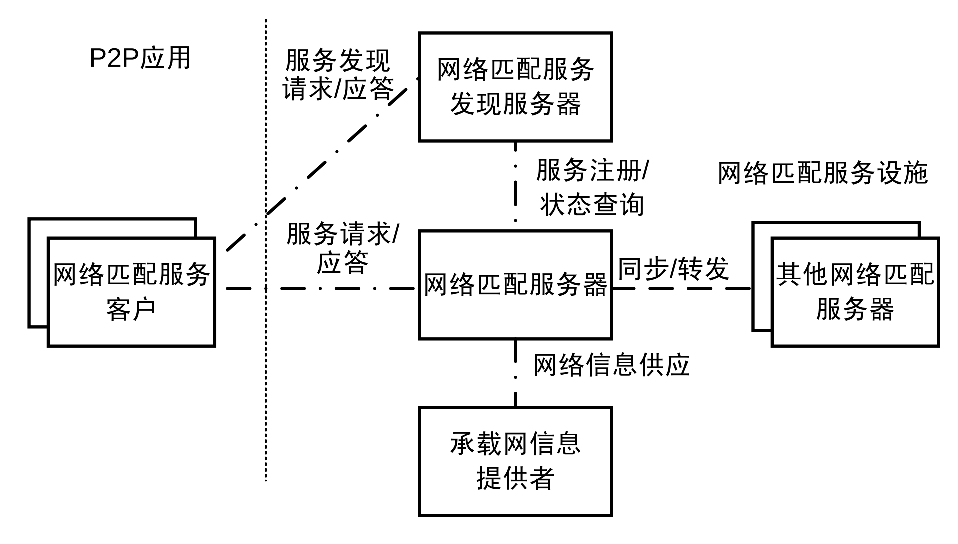


图5. 网络匹配服务的基本架构

4.2 位置感知的拓扑构建

获取位置信息的主要目的是为了构建位置感知的覆盖拓扑。在存在 Tracker 的系统中(如 BT 和 PPLive)，Tracker 维护了参与当前会话的节点。当一个节点尝试加入或希望进行邻居替换时，会向 Tracker 请求候选邻居集合。Tracker 可以根据位置信息为请求节点分配一组位置相近的候选节点。对等传输应用节点也可在运行过程中依据探测的信息自主地优化自身的拓扑。运营商可以通过部署 Proxy-Tracker 的方式来辅助应用节点建立位置感知的拓扑。Proxy-Tracker 部署在一个网络的边界，维护本网络参与某个对等传输会话的节点信息。Proxy-Tracker 截获对等传输应用的 Tracker 发送给节点的候选邻居响应消息，修改候选邻居集合，将一些距请求节点位置较远的节点替换成与请求节点位于同一网络内的节点，从而辅助请求节点建立位置感知的拓扑。网络运营商还可以通过提供重定向服务器实现位置感知的拓扑构建。重定向服务器维护本域内的对等传输应用节点信息和拥有的对象信息，截获用户的连接或数据请求，并重定向连接或数据请求。

5 对等传输数据调度算法

数据调度算法是优化对等传输流量的另一选择。但针对数据调度算法的研究通常是从应用层性能提升的角度出发的，没有考虑对底层网络运营商的影响。数据调度算法对第三代的对等传输内容分发技术（即原文件被划分成多个数据块，节点构成协作网络来完成原文件的分发）有效。这类系统中，节点通常只基于邻居节点的局部信息（即自身和邻居的数据块信息）来进行决策，确定需要下载哪个数据块。典型的数据调度算法包括：随机调度、局部最稀缺优先（local rarest first）调度和最近新兴的基于网络编码的调度[16][17]。

单靠数据调度算法无法优化对等传输流量，必须结合位置感知的下载策略才能实现优化流量的目标。图 6 说明了三种不同的数据调度算法对对等传输流量的影响。假设初始时刻只有节点 A 拥有数据块 a、b、c、d。图 6(a)采用随机下载决策，节点 B、C、D、E 可能同时向节点 A 请求同一个数据块，如数据块 a。这将导致 B、C、D、E 之间的连接在接下来的数据传输周期内无法被利用。图 6(b)采用局部最稀缺优先策略进行下载决策，但即使如此，

也会导致数据分布的不均匀。例如，节点 B 先向节点 A 请求数据块 a，然后节点 C 向节点 A 请求数据块，根据局部最稀缺优先策略，则节点 C 可以向节点 A 请求数据块 b，然后节点 D 向 A 请求数据。依据该策略，节点 D 可能会向节点 A 请求数据块 a，同样，节点 E 可能向节点 A 请求数据块 b。这样，B、C、D、E 只拥有数据块 a 和 b，在经过另一个传输周期后，它们之间的域内链路将无法利用，新数据只能从节点 A 获取。图 6(c)是采用网络编码的情况，节点 A 分别向节点 B、C、D、E 发送随机线性编码的编码数据块。当有限域足够大时，如 F(28)或 F(216)，即可保证传输给节点 B、C、D、E 的四个数据块以很高的概率线性无关。这样，节点 B、C、D、E 可以通过域内链路完成数据的传输从而解码出原始数据块，实现最优化的资源利用。

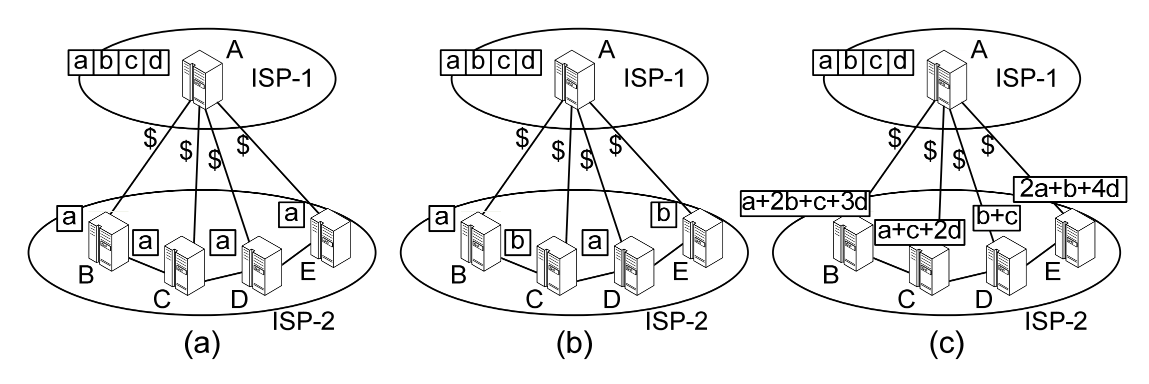


图6. 数据调度算法对对等传输流量优化的作用
(a)随机调度策略；(b)局部最稀缺优先策略；(c)网络编码

研究表明[15]，在同样的位置感知能力支持下，基于局部最稀缺优先的数据调度产生的域间流量冗余度能比基于随机数据调度产生的域间流量冗余度降低 80%~90%。但即使是位置感知能力最强的设置，基于局部最稀缺优先的数据调度的域间流量冗余度依然在 3.0 以上，即平均每个数据块要进入一个域 3 次以上。我们在同样的位置感知能力支持下，对比了基于局部最稀缺优先的数据调度和基于网络编码的数据调度在对等传输流量优化上的作用。我们的实验以中国 AS 拓扑的 4-core 作为底层的 AS 拓扑，系统包含 1000 个节点，节点平均度为 10。实验结果如图 7 所示，当位置感知能力增强时，两种数据调度算法产生的域间流量都出现了明显的下降。但当位置感知能力并不太弱时，网络编码所导致的域间流量冗余度低于局部最稀缺优先调度算法的一半。

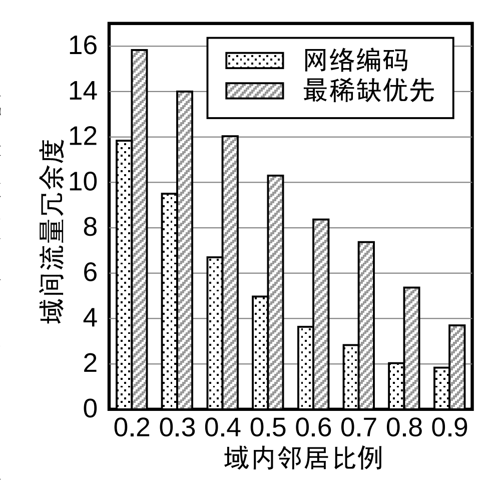


图7. 网络编码和局部最稀缺优先数据调度策略产生的域间数据冗余度

6 比较、总结和展望

6.1 比较

对等传输缓存、位置感知技术和数据调度算法虽然都可以用于优化对等传输流量，但是三者在适用性、独立性和技术的实现方法上有所区别。表 2 总结了三种技术在适用性、独立性和实现方的区别。

表2. 不同的对等传输流量优化技术在适用性、可独立使用性和实现方的区别

	对等传输缓存	位置感知技术	数据调度
适用性	适用于所有对等传输应用；但随着对等传输网络位置感知能力的增强，有效性将降低	适用于所有对等传输应用	仅适用于 BT 和 PPLive 等存在合作下载网络的应用
独立性	可独立使用	可独立使用	需要和位置感知技术结合使用
实现方	网络运营商	网络运营商、内容提供商或第三方	内容提供商

6.2 总结和展望

对等传输流量优化是对等传输应用得以健康、良性和可持续发展的前提。本文从对等传输缓存技术、位置感知技术和数据调度算法三个角度综述了不同的技术对优化对等传输流量的作用。

虽然对等传输流量优化目前在理论上的研究已经取得了长足的进展，但是被网络运营商和内容提供商的广泛采用还要经过一段努力。一方面，许多理论技术的有效性仅通过仿真试验得到验证，缺乏在互联网上大规模的实际测量结果。真正的互联网环境下，网络地址转换（Network Address Translation，NAT）和防火墙的存在可能会大幅降低位置感知能力的效果。因而，在实际的互联网上进行大规模的测试评估成为验证位置感知技术对对等传输流量优化有效性程度的迫切任务。另一方面，在技术可行的前提下，如何建立适合网络运营商、内容提供商、终端用户之间的商业模型是推动该技术的关键。

参考文献：

[1] H. Y. Xie, Y. R. Yang, A. Krishnamurthy, Y. B. Liu, and A. Silberschatz, “P4P: Provider portal for applications”, *ACM SIGCOMM* 2008.

[2] A. Wierzbicki, N. Leibowitz, M. Ripeanu, and R. Wozniak, “Cache replacement policies revisited: the case of P2P traffic”, in *Proc. 4th Int. Workshop on Global and Peer-to-Peer Computing (GP2P’04)*, Chicago, IL, Apr. 2004, pp.182-189.

[3] M. Hefeeda and O. Saleh, “Traffic modeling and proportional partial caching for peer-to-peer systems”, *IEEE/ACM Trans. Networking*, vol.16, no.6, pp.1447-1460, Dec, 2008

[4] 邹嵘，“基于 P2P Cache 的 P2P 流量优化技术”，*电信网技术*，2009 年 1 月

[5] M. Ripeanu, L. Foster and A. Iamnitchi, “Mapping the Gnutella Network: Properties of Large-scale Peer-to-Peer Systems and Implications for System Design”, *IEEE International Computing Journal Special Issue on peer-to-peer networking*, 2002.

[6] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, “IDMaps: A Global Internet Host Distance Estimation Service”, *IEEE/ACM Transactions on Networking*, 9(5): 525-540, 2001.

[7] Sylvia Ratnasamy, Mark Handley, Richard Karp and Scott Shenker, “Topologically-aware Overlay construction and server selection”, In *Proc. INFOCOM 2002*, 2002

[8] K. P. Gummadi, S. Saroiu, and S. D. Gribble, “King: Estimating Latency between Arbitrary Internet End Hosts”, *IMC’02*, 2002.

[9] D. Leonard and D. Loguinov, “Turbo King: Framework for Large-Scale Internet Delay Measurements”, *INFOCOM’08*, 2008.

[10] D. R. Choffnes, and F. E. Bustamante, “Taming the Torrent: A Practical Approach to Reducing Cross-ISP Traffic in Peer-to-Peer Systems”, *SIGCOMM’08*, 2008.

[11] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, “Vivaldi: A decentralized network coordinate system”, *in Proc. ACM SIGCOMM*, *Portland*, *OR*, Aug, 2004.

[12] E.Ng and H. Zhang, “Predicting Internet network distance with coordinates-based approaches”, in *Proc IEEE INFOCOM*, 2002

[13] C. Cramer, K. Kutzner and T. Fuhrmann, “Bootstrapping locality-aware P2P networks”, in.*Proc. 12th Int. Conf. Networks(ICON’04)*, vol. 1, pp. 357-361, 2004.

[14] 中国科学院计算所等，“基于承载网感知的 P2P 流量优化技术框架”，*通信行业标准*，报批稿。

[15] Bindal et al, “Improving traffic locality in BitTorrent via Biased neighbor selection”, in *Proc. IEEE ICDCS*, 2006

[16] R. Ahlswede, N. Cai, S. R. Li, and R. W. Yeung, “Network Information flow”, *IEEE Trans. Information Theory*, 2000

[17] S. R. Li, R. W. Yeung, and N. Cai, “Linear network coding”, *IEEE Trans. Information Theory*, vol. 49, no. 2, 2003

[18] 唐明董，张国清，杨景，傅川，廖祝华，“P2P 流量优化技术综述”，*电信网技术*，2009 年 1 月

作者简介：

张国强 中国科学院计算技术研究所助理研究员，博士，Email: guoqiang@ict.ac.cn

唐明董 中国科学院计算技术研究所博士生，湖南科技大学计算机学院讲师

程苏琦 中国科学院计算技术研究所

张国清 中国科学院计算技术研究所硕士生导师，博士，Email: gqzhang@ict.ac.cn